X-547-64-295

NASA TM X-55140

LOCAL ERROR CONTROL IN NUMERICAL INTEGRATION THROUGH OPTIMIZING THE ORDER OF THE INTEGRATOR

OCTOBER 1964

GODDARD SPACE FLIGHT CENTER

GREENBELT, MARYLAND

LOCAL ERROR CONTROL IN NUMERICAL INTEGRATION THROUGH OPTIMIZING THE ORDER OF THE INTEGRATOR

by

W. S. Soar

CONTENTS

LOCAL ERROR CONTROL IN NUMERICAL INTEGRATION THROUGH OPTIMIZING THE ORDER OF THE INTEGRATOR

THE NATURE OF NUMERICAL INTEGRATION

Numerical integration consists of replacing the function to be integrated by another of simpler form, usually a polynomial, which is then integrated by some standard formula. This is true whether the integrand is a function of the independent variable only, as in the case

$$y' = f(x),$$

or is a function of both independent and dependent variables, as in the differential equation

$$y' = f(x, y).$$

The first case is the problem of the integral calculus, and is solved by numerical quadrature. Here, for a given x we can evaluate y', and can substitute values of y', or finite differences of y', into appropriate formulas to obtain the integral y over some domain.

In the second case we require knowledge of the integral y for a given x before we can evaluate y'. This is like having to know the answer before we can ask the question. A common approach to this problem is to estimate y at the specified x, substitute it into the differential equation to obtain an estimate of y', after which y is computed as in ordinary quadrature. This procedure is called the predictor and corrector method, of which numerous forms exist. If the first computed y is not good enough, it may be substituted into the differential equation exactly as was the original estimate of y. An improved y' is then found, which produces a better y.

We shall assume throughout that y' is approximated by a polynomial. Then the success of the numerical integration depends upon how well y' is represented by the polynomial which replaces it, for the formulas integrate the polynomial exactly.

HAZARDS OF POLYNOMIAL APPROXIMATION

It is well known that polynomials, under appropriate circumstances, can be made to approximate almost any function to arbitrary accuracy. The Weierstrass approximation theorem assures us that if $f(x)$ is a real function continuous on the closed interval $[a, b]$, then for every positive number ϵ there exists a polynomial $P(x)$ of degree n,

$$P(x) \equiv \sum_{k=0}^{n} a_k x^k,$$

such that $|f(x) - P(x)| < \epsilon$ for all x in $[a, b]$.

But this theorem is widely misinterpreted. To begin with, the fact that the desired polynomial exists does not imply that we can find it. That does not suggest we lack skills or methods. We may be unable to find the polynomial of arbitrary accuracy simply because we cannot get enough information about it.

Of one thing we may be sure. The existence of the desired polynomial does not mean it is of low degree. But here is something often misunderstood: A polynomial of higher degree does not necessarily give a better approximation.

Consider the example

$$f(x) = \frac{1}{1 + x^2}.$$

Values of $f(x)$ are given for equal intervals in x.

Table 1

x	-3	-2	-1	0	1	2	3	4
f(x)	.1	.2	.5	1	.5	.2	.1	.05882 352

We now interpolate among these points to find the functional value $f(.5)$, which is the same thing as saying we approximate the function by a polynomial which we evaluate for $x = .5$. Since $f(x)$ is of simple form, we can evaluate $f(.5)$ and know the correct answer in advance.

$$f(.5) = \frac{1}{1 + (.5)^2} = .8$$

To represent $f(x)$ by a polynomial of degree n we pass the polynomial exactly through $n + 1$ known values of $f(x)$. The Weierstrass theorem assures that a polynomial exists which approximates $f(x)$ uniformly over the domain in question, but let us be realistic and compute our polynomial from information obtained as close as possible to the one value we are interested in. Thus, in the table below, when $n + 1$ is even the points are symmetric with respect to $x = .5$; when $n + 1$ is odd one extra point is taken before $x = .5$.

Consider the table, then, which gives the values of $P(.5)$ where $P(x)$ is a polynomial of the degree specified.

Table 2

Degree n of Polynomial	Placement of Points Before	After	P(.5)	See Text
1	1	1	.75	.75
2	2	1	.875	.80961 540
3	2	2	.8	.80192 309
4	3	2	.85625	.80018 986
5	3	3	.81406 25	.80008 631
6	4	3	.85015 625	.80000 002
7	4	4	.82136 030	.80000 156

Contrary to common belief, the attempt to approximate $f(x)$ by polynomials of higher degree leads to failure. Why do not those of degree greater than 3 do as well as the cubic? Does this example disprove Weierstrass?

The difficulty, aside from the fact that a rational function is not easily fitted by a polynomial, is that no attention was given to the interval $[a, b]$ over which the function was to be approximated. For the polynomial of degree one the interval used was $[0, 1]$; for degree 2 the interval was $[-1, 1]$; for degree 7 the interval was $[-3, 4]$. A final column is added to the table showing interpolated values of $f(.5)$ when all information consists of $n + 1$ points equally spaced within the interval $[0, 1]$.

It is clear, then, that increasing the degree of the approximating polynomial does not always guarantee a better approximation. It should be clear also that changing to an integration formula of higher order does not necessarily assure a more accurate integral.

CONTROLS IN NUMERICAL INTEGRATION

If we seek to increase the accuracy of the integral

$$\int_a^b f(x)\,dx$$

by using a formula of higher order, then, in general, we require more information about $f(x)$ in $[a, b]$, which is to say the integration interval $\Delta x = h$ from one point to the next must be shortened.

And here we have two important controls to govern the accuracy of numerical integration:

1. We may select the order of the formula.
2. We may specify the step size h.

They are intimately related; and both are related to the third control:

3. We may decide the number of significant digits to be retained in the computation. (This last decision usually consists of directing the use of single or double precision mode of machine operation).

Very well, you may say, but how do we apply the controls?

Ordinarily when one undertakes a job of numerical integration his first choice is the integration formula. He can decide later what step size to use, and whether to run in single or double precision. But how does he pick the formula?

FROM INTERPOLATION TO INTEGRATION

First let us consider what an integration formula is and how we get it. Although it is not the only way, one common way to obtain an integration formula is to integrate formally an interpolation formula. The interpolation formula is

usually an equation into which we may substitute either known values of the given function or their finite differences, together with an arbitrary value of the independent variable for which we wish to know the corresponding value of the function. The formula automatically fits a polynomial exactly to the given points and evaluates it at the arbitrary point without our ever knowing explicitly what the polynomial is.

If we did know the interpolating polynomial, we could integrate it term by term between selected limits and obtain an approximation to the definite integral of the given function between those limits. But by integrating the interpolation formula we obtain an integration formula which is general. Suppose we choose the Lagrange interpolation formula, which employs functional values directly: we get an integration formula in terms of functional values. Suppose we choose Stirling's formula, which employs central differences of the functional values: we get an integration formula in terms of central differences, as we shall see later.

In general the integration formulas expressed in differences are no better or worse than those expressed in ordinates. Often one form is readily derived from the other. For certain purposes, however, one may be better suited. For our immediate purpose of selecting the order of formula to do a specific job, we find an advantage in the formulas expressed in differences.

DERIVATION OF THE "COWELL'S" FORMULA

Suppose we are called upon to integrate the equations of motion of an artificial earth satellite, in rectangular coordinates. We have here a system of three second-order equations which must be integrated simultaneously. They are of the form

$$\ddot{x} = f(x, y, z; t),$$

with corresponding equations for y and z. The first derivative $\dot{x}$ does not appear explicitly.

We digress briefly to consider the derivation of Taylor's formula with remainder, which we shall use in obtaining the integration formula.

Making all necessary assumptions about the continuity of $f(x)$ and the existence and continuity of its derivatives of all orders, we can construct the formula by integrating the n-th derivative n times from x_0 to x.

$$\int_{x_0}^{x} f^{(n)}(x)\,dx = f^{(n-1)}(x) - f^{(n-1)}(x_0)$$

$$\int_{x_0}^{x}\int_{x_0}^{x} f^{(n)}(x)\,dx\,dx = \int_{x_0}^{x} f^{(n-1)}(x)\,dx - \int_{x_0}^{x} f^{(n-1)}(x_0)\,dx$$

Since the integrand $f^{(n-1)}(x_0)$ is a constant, we have

$$\int_{x_0}^{x}\int_{x_0}^{x} f^{(n)}(x)\,dx\,dx = f^{(n-2)}(x)\Big]_{x_0}^{x} - (x - x_0)\,f^{(n-1)}(x_0)$$

$$= f^{(n-2)}(x) - f^{(n-2)}(x_0) - (x - x_0)\,f^{(n-1)}(x_0).$$

This equation can be rewritten

$$f^{(n-2)}(x) = f^{(n-2)}(x_0) + (x - x_0)\,f^{(n-1)}(x_0) + \int_{x_0}^{x}\int_{x_0}^{x} f^{(n)}(x)\,dx\,dx.$$

Although we have performed only two integrations, already the Taylor's formula has begun to take shape. Each additional integration will provide another term in the series. But this is far enough. Let $n = 2$. The zero-th derivative of $f(x)$ is simply $f(x)$, and the last equation becomes

$$f(x) = f(x_0) + (x - x_0)\,f'(x_0) + \int_{x_0}^{x}\int_{x_0}^{x} f''(x)\,dx\,dx. \qquad (1)$$

Using the notation $y = f(x)$, we evaluate $y_1 = f(x_1)$ and $y_{-1} = f(x_{-1})$ by means of (1).

$$y_1 = y_0 + hy_0' + \int_{x_0}^{x_1} \int_{x_0}^{x} y'' \, dxdx$$

$$y_{-1} = y_0 - hy_0' + \int_{x_0}^{x_{-1}} \int_{x_0}^{x} y'' \, dxdx$$

Add these equations and obtain

$$y_1 - 2y_0 + y_{-1} = \int_{x_0}^{x_1} \int_{x_0}^{x} y'' \, dxdx + \int_{x_0}^{x_{-1}} \int_{x_0}^{x} y'' \, dxdx. \qquad (2)$$

It will be observed that the left side of this equation is $\delta^2 y_0$, the second central difference in Sheppard's notation. Further, the equation does not contain the first derivative, a significant fact since we are seeking a method for integrating second-order differential equations where the first derivative is absent.

To dispose of the two integrals on the right side of this equation we make use of Stirling's interpolation formula, which is expressed in central difference notation.

$$y_s = y_0 + \sum_{n=1}^{\infty} \left\{ \mu\delta^{2n-1} y_0 \cdot \frac{s(s^2 - 1^2)(s^2 - 2^2) \cdots \left[s^2 - (n-1)^2\right]}{(2n-1)!} \right.$$

$$\left. + \delta^{2n} y_0 \cdot \frac{s^2(s^2 - 1^2)(s^2 - 2^2) \cdots \left[s^2 - (n-1)^2\right]}{(2n)!} \right\} \qquad (3)$$

Here

$$\delta y_0 = y_{\frac{1}{2}} - y_{-\frac{1}{2}}$$

$$\mu y_0 = \frac{1}{2}\left(y_{\frac{1}{2}} + y_{-\frac{1}{2}}\right)$$

$$s = \frac{x - x_0}{h}$$

where h is the length of equal intervals in x.

It is clear that the factors

$$\frac{\mu\delta^{2n-1}}{(2n-1)!}\, y_0$$

and

$$\frac{\delta^{2n}}{(2n)!}\, y_0$$

in (3) are constants, so that the formula can be integrated with respect to the variable s. Let us, then, represent y″ in (2) by Stirling's interpolation formula, and integrate twice as (2) requires.

Stirling's formula is an infinite power series. If we expand it to some point and truncate the series, we have simply a polynomial, which we can integrate term by term.

Upon carrying out the integration and adding the two integrals as called for in (2), we find that all terms in $\mu\delta^{2n-1}\, y_0''$ vanish; and, after considerable arithmetic involving the addition of common fractions, which are kept as common fractions for exactness, we finally obtain

$$\delta^2 y_n = h^2\left(y_n'' + \frac{1}{12}\,\delta^2 y_n'' - \frac{1}{240}\,\delta^4 y_n'' + \frac{31}{60480}\,\delta^6 y_n'' - \frac{289}{36\ 28800}\,\delta^8 y_n'' + \cdots\right). \quad (4)$$

This is the formula used by Cowell and Crommelin to integrate the orbit of Halley's comet. Nowadays it is most often used in a modified form with values of y'' instead of central differences of y'', but for our purpose the difference form has special advantages, as was suggested earlier.

The first advantage is that (4) itself is an infinite series. How many terms appear depends upon how many were retained and integrated in Stirling's formula. But the important thing is that if we wish to use a formula of order higher than (4), we simply compute the coefficient of the next-higher term without changing the lower-order terms. We shall see later that it has another advantage.

To use (4) as written, we compute a central difference table to ninth differences $\delta^9 y''_n$. Extrapolate values of $\delta^9 y''$ beyond $\delta^9 y''_n$, assuming that $\delta^9 y''_{n+i} = \delta^9 y''_n$, and sum across the table until we have the necessary values to compute $\delta^2 y_n$ from the formula. Since

$$\delta^2 y_n = y_{n+1} - 2y_n + y_{n-1} ,$$

we have

$$y_{n+1} = \delta^2 y_n + 2y_n - y_{n-1} .$$

Here, then, is an estimate of the integral y at the $n+1$-th step, from which we obtain an estimate of y''_{n+1} by substituting into the differential equation. With this y''_{n+1} compute the line of differences out to $\delta^9 y''_{n+1}$, which we had formerly estimated by extrapolation. With these corrected differences, use the formula to find a better $\delta^2 y_n$. Compute a better y_{n+1}. Iterate until y_{n+1} converges.

USE OF DIFFERENCE TABLE IN OPTIMIZING CONTROLS

Take a good look at (4). Suppose we have the indicated differences and a value for h. Then we can compute the contribution of each term, and can see readily whether the kth term affects the last significant digit retained in the computation. Here all three controls come together. For a given h we may see that $h^2 a_k \delta^k y''_n$ is less than 10^{-8} and may be omitted in single precision operation, but is greater than 10^{-15} and must be included for double precision.

We can control the magnitude of $h^2 a_k \delta^k y''_n$ in two ways: by increasing k or by decreasing h. In (4) we see that coefficients a_k of successive terms grow smaller, and it may be shown that this trend would continue if more terms were

computed. It is also true in general--although not always, and the exceptions constitute a dangerous trap*--that higher-order differences tend to zero. Hence, by taking more terms we make both of the last two factors, and their product, smaller.

Suppose, on the other hand, that we take a shorter step size h, say cut h in half. Immediately we multiply each term by one fourth because of the h^2 factor; but, beyond this, the differences also are reduced. First differences are approximately half as large as before; second differences, one fourth as large; nth differences, $1/2^n$ as large. Therefore, cutting h makes the first and third factors, and their product, smaller.

All this would be nice if we could obtain the difference table over the domain of integration. We could find an optimum combination of h and k, step size and order of formula, to squeeze out the last bit of significance at each integration step.

In the case of simple quadrature this actually can be done if the quadrature formula is expressed in differences. But for differential equations we are not likely ever to have a difference table of the actual quantities we wish to integrate. We obtain estimates of these quantities only by substituting into the differential equation the estimates of the integral at successive steps.

Of course even here we can look at the successive terms of the formula and see which actually are contributing to the solution, and how much. But this study can be made only after we have picked some formula, and some h, and some number of digits to carry.

In the problem we assigned ourselves earlier, that of integrating the equations of motion of an artificial earth satellite, we can do better. We can form a difference table from values of acceleration based upon the two-body solution. Admittedly, the perturbed orbit is not the unperturbed orbit; but in most cases the greatest perturbing force, that due to the nonsphericity of the earth, is less than 10^{-3} times the central attracting force. By observing how the unperturbed values of y'' change with time we can get some notion of how the actual values change.

If we consider only the order of the integration formula, we probably should assume that a rapidly changing integrand will require a formula of higher order than a slowly changing one. It might be wise, then, to look at the table of differences in those parts of the orbit where accelerations are changing most rapidly as well as where they are changing slowly.

*Form differences of the values in Table 1. The sixth difference is −7.2; the seventh, 11.8 +.

We can find the values of true anomaly corresponding to the maximum and minimum rates of change of acceleration as a function eccentricity. But this really is not much help because we are integrating not the total acceleration but its rectangular components, and where one component is changing rapidly another may be changing slowly. It would seem wise, therefore, to pick some likely value of h, say one minute, and compute the difference table for two-body accelerations from perigee to apogee.

Table 3A corresponds to a 20-step section in the orbit of an actual satellite, but without perturbations. The quantities labelled "Maximum in Block" are the largest among all the kth differences for the component specified. Thus for $\ddot{x}$ there were 14 values of 6th differences, of which the largest was $.12600 \times 10^{-3}$. Here h is one minute of time, or .07436 492 canonical time units.

During the part of the orbit covered by this table the magnitude of the x-component of position reached a maximum of 1.09663+ canonical units of length. In single precision this is represented as $.10966389 \times 10^{+1}$; in double precision, as $.10966389829835 \times 10^{+1}$. The last three columns of the table are kth terms in the formula (4). Clearly terms in $\delta^6 \ddot{x}$ and higher are not significant in single precision operation: they could add nothing to x as represented above. But all terms listed are significant for double precision.

Perhaps a term of order higher than $\delta^{12} \ddot{x}$ might also be significant. Where do we stop? We have seen integration formulas which actually used the equivalent of twelfth differences in (4). But at some point it would seem reasonable to cut the step size and use fewer differences, *i.e.*, a formula of lower order.

Table 3B shows what happens in the same part of the orbit when h is cut to half a minute, or .03718 246 canonical time units. Already the terms in fourth differences are insignificant for single precision, while those in eighth differences barely affect the last digit in double precision.

It is unnecessary, or course, to form the products as given in Table 3. Instead, we can consider the first two factors only, and decide how large the kth difference must be for significance. In double precision, for example, if x is approximately one canonical unit of length, as above, the 15th significant digit is of the order of 10^{-14}, and the product of the three factors must be no less than this quantity. Thus, for h = 1 minute, $\delta^{12} \ddot{x}$ must be no less than $.27805 \times 10^{-5}$.

$$h^2 a_{12} \delta^{12} \ddot{x} = (.13825 \times 10^{-2})(.26014 \times 10^{-5}) \delta^{12} \ddot{x} \geq 10^{-14}$$

$$(.35964 \times 10^{-8}) \delta^{12} \ddot{x} \geq 10^{-14}$$

$$\delta^{12} \ddot{x} \geq .27805 \times 10^{-5}.$$

Table 3A

h = 1 min. $\simeq$ .07436 492 canonical time units

$h^2 = .55301 \times 10^{-2}$

k	$\lvert a_k \rvert$	Maximum in Block			$h^2 a_k \delta^k \ddot{x}$	$h^2 a_k \delta^k \ddot{y}$	$h^2 a_k \delta^k \ddot{z}$
		$\lvert \delta^k \ddot{x} \rvert$	$\lvert \delta^k \ddot{y} \rvert$	$\lvert \delta^k \ddot{z} \rvert$			
2	$.83333 \times 10^{-1}$	$.16354 \times 10^{-1}$	$.10209 \times 10^{-1}$	$.47243 \times 10^{-2}$	$.75366 \times 10^{-5}$	$.47047 \times 10^{-5}$	$.21771 \times 10^{-5}$
4	$.41666 \times 10^{-2}$	$.71380 \times 10^{-3}$	$.98164 \times 10^{-3}$	$.51168 \times 10^{-3}$	$.16447 \times 10^{-7}$	$.22619 \times 10^{-7}$	$.11790 \times 10^{-7}$
6	$.51257 \times 10^{-3}$	$.12600 \times 10^{-3}$	$.14600 \times 10^{-3}$	$.10473 \times 10^{-3}$	$.35715 \times 10^{-9}$	$.41385 \times 10^{-9}$	$.29686 \times 10^{-9}$
8	$.79641 \times 10^{-4}$	$.44652 \times 10^{-4}$	$.22886 \times 10^{-4}$	$.10397 \times 10^{-4}$	$.19666 \times 10^{-10}$	$.10080 \times 10^{-10}$	$.45791 \times 10^{-11}$
10	$.13898 \times 10^{-4}$	$.67087 \times 10^{-5}$	$.12880 \times 10^{-4}$	$.66286 \times 10^{-5}$	$.51561 \times 10^{-12}$	$.98992 \times 10^{-12}$	$.50945 \times 10^{-12}$
12	$.26014 \times 10^{-5}$	$.56675 \times 10^{-5}$	$.22767 \times 10^{-5}$	$.18297 \times 10^{-5}$	$.81533 \times 10^{-13}$	$.32753 \times 10^{-13}$	$.26322 \times 10^{-13}$

Table 3B

$h = 1/2$ min. $\cong$ .03718 246 canonical time units

$h^2 = .13825 \times 10^{-2}$

k	$\lvert a_k \rvert$	Maximum in Block			$h^2 a_k \delta^k \ddot{x}$	$h^2 a_k \delta^k \ddot{y}$	$h^2 a_k \delta^k \ddot{z}$
		$\lvert \delta^k \ddot{x} \rvert$	$\lvert \delta^k \ddot{y} \rvert$	$\lvert \delta^k \ddot{z} \rvert$			
2	$.83333 \times 10^{-1}$	$.90116 \times 10^{-3}$	$.12224 \times 10^{-2}$	$.93323 \times 10^{-3}$	$.10382 \times 10^{-6}$	$.14083 \times 10^{-6}$	$.10752 \times 10^{-6}$
4	$.41666 \times 10^{-2}$	$.17614 \times 10^{-4}$	$.96060 \times 10^{-5}$	$.31525 \times 10^{-5}$	$.10146 \times 10^{-9}$	$.55333 \times 10^{-10}$	$.18159 \times 10^{-10}$
6	$.51257 \times 10^{-3}$	$.23060 \times 10^{-6}$	$.25858 \times 10^{-6}$	$.25020 \times 10^{-6}$	$.16341 \times 10^{-12}$	$.18324 \times 10^{-12}$	$.17730 \times 10^{-12}$
8	$.79641 \times 10^{-4}$	$.15199 \times 10^{-7}$	$.10168 \times 10^{-7}$	$.34307 \times 10^{-8}$	$.16735 \times 10^{-14}$	$.11195 \times 10^{-14}$	$.37773 \times 10^{-15}$
10	$.13898 \times 10^{-14}$	$.66474 \times 10^{-9}$	$.59098 \times 10^{-9}$	$.54233 \times 10^{-9}$	$.12772 \times 10^{-16}$	$.11355 \times 10^{-16}$	$.10420 \times 10^{-16}$
12	$.26014 \times 10^{-5}$	$.44138 \times 10^{-10}$	$.61070 \times 10^{-10}$	$.22333 \times 10^{-10}$	$.15874 \times 10^{-18}$	$.21963 \times 10^{-18}$	$.80319 \times 10^{-19}$

AN AUTOMATIC FORMULA SELECTOR

We have known all along that a computer can do anything we can tell it how to do. It seems not unreasonable to suppose we might tell it how to select its own integration formula in a particular instance.

Suppose we wish to integrate an orbit of such and such properties. The computer might integrate one revolution with a formula of high order and short step size. Then, by examining the differences of various orders, it could decide that terms in the formula beyond a certain k add nothing to the solution. On the other hand, it might decide that because the k it selected is smaller than some predetermined value, the step size should be lengthened.

To be sure, the differences would not be differences of the actual accelerations we want to integrate, but they would be differences of the only accelerations we ever can integrate. And they would include the perturbing forces.

Having selected the order of the formula and the step size, the actual computation might then be carried out using an equivalent formula in terms of the accelerations directly instead of differences of accelerations, thereby avoiding further computation and storing of the difference table.

The orbital parameters which govern what order of formula should be used, and what step size, are the semimajor axis a and the eccentricity e. It is reasonable to think that if some study is given to the results of the automatic formula selection, what formula goes with which a and e, we might then change the integration program so that it makes its selection on the basis of the initial input alone.

Whether it is desirable to change the formula or step size within one revolution of the satellite would certainly depend upon a and e. For any a, but $e = 0$, there seems little need for changing either the formula or the step size. For large a and large e, however, it might be wise to change either formula or step size, or both, within one revolution. Again, since the rate of change of acceleration depends upon these two parameters, the computer could decide, on the basis of the initial conditions, whether to make changes and, if so, where.